ARBITRARY STATE MARKOVIAN DECISION PROCESSES

Sheldon M Ross

Stanford University
Stanford, California

January 1968

ARBITRARY STATE MARKOVIAN DECISION PROCESSES

by

Sheldon M. Ross

TECHNICAL REPORT NO 105

January 8, 1968

Supported by the Army, Navy, Air Force, and NASA under

Contract Nonr-225(53)(NR-042-002)

with the Office of Naval Research

Gerald J. Lieberman, Project Director

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

NONTECHNICAL SUMMARY

A Markovian Decision Process is a process which is observed at distinct time points to be in some state. After observing the state of the system an action is chosen - corresponding to the action (and the present state) a cost is incurred and the transition probabilities for the next state are determined. A policy is any rule for choosing actions. Corresponding to each policy there is an expected long run average cost per unit time. This paper is concerned with finding an optimal policy - i.e. one whose associated average cost is minimal.

For example we might have a tool which wears out with time. The state of the system could be the length of the tool, and the possible actions could be either to replace the tool or not Associated with each state there would be an operating cost. Thus a policy is a rule for determining when to replace the tool and an optimal one is one which minimizes the long run average cost.

In the past most of the work in this area has been done under the assumption that the state space is countable. In this paper we let the state space be arbitrary. For example, in the tool problem given above it is natural to let the state space be the continuum of possible values of the length of the tool

This paper presents sufficient conditions for the existence of an optimal policy and for it to be of simple type This type - called stationary deterministic - is of the form of a function mapping the state space into the action space For example, in the tool problem

a stationary deterministic policy would replace whenever the length of the tool is in some specified set of real numbers. The method employed is to treat the average cost problem as a limit of either the discounted cost problem or the nondiscounted n-stage problem. We also show how, in a special case, the average cost problem may be reduced to a discounted cost problem.

ARBITRARY STATE MARKOVIAN DECISION PROCESSES

Sheldon M. Ross

1. Introduction

We are concerned with a process which is observed at times $t = 0,1,2,\ldots$ and classified into one of a possible number of states. We let $\mathcal{X}$ denote the state space of the process. $\mathcal{X}$ is assumed to be a Borel subset of a complete separable metric space, and we let $\mathcal{B}$ be the σ-algebra of Borel subsets of $\mathcal{X}$. After each classification an action must be chosen and we let A, assumed finite, denote the set of all possible actions.

Let $\{X_t;\ t = 0,1,2,\ldots\}$ and $\{\Delta_t;\ t = 0,1,2,\ldots\}$ denote the sequence of states and actions; and let $S_{t-1} = (X_0,\Delta_0,\ldots X_{t-1},\Delta_{t-1})$. It is assumed that for every $x\epsilon\mathcal{X}$, $k\epsilon A$ there is a known probability measure $P(\cdot|x,k)$ on $\mathcal{B}$ such that, for some version, $P\{X_{t+1}\epsilon B|X_t = x,\ \Delta_t = k,\ S_{t-1}\} = P(B|x,k)$ for every $B\epsilon\mathcal{B}$, and all histories S_{t-1}. It is also assumed that for every $k\epsilon A$, $B\epsilon\mathcal{B}$, $P(B|\cdot,k)$ is a Baire function on $\mathcal{X}$.

Whenever the process is in state x and action k is chosen then a bounded (expected) cost $C(x,k)$ - assumed, for fixed k, to be a Baire function in x- is incurred.

A policy R is a set of Baire functions $\{D_k(S_{t-1},x)\}_{k\epsilon A}$ satisfying $D_k(S_{t-1},x) \geq 0$ for all $k\epsilon A$, and $\sum_{k\epsilon A} D_k(S_{t-1},x) = 1$ for every (S_{t-1},x). The interpretation being: if at time t the history S_{t-1} has been observed and $X_t = x$ then action k is chosen with probability $D_k(S_{t-1},x)$. R is said to be stationary if $D_k(S_{t-1},x) = D_k(x)$ for every S_{t-1}; R is said to be stationary deterministic if $D_k(x)$ equals 0 or 1 for all x,k.

For any policy R, let $\phi(x,R) = \limsup_{n \to \infty} 1/n \sum_{t=0}^{n} E_R [C(X_t,\Delta_t)|X_0=X]$. Thus $\phi(x,R)$ is the expected average cost per unit time when the process starts in state x and policy R is used.

In [4] , under the assumption that χ is denumerable, a number of results dealing with the average cost criterion were proven. The method employed was to treat the average cost problem as a limit (as the discount factor approaches unity) of the discounted cost problem. In this paper we generalize some of these results to arbitrary state spaces. We also show how to treat the average cost problem as a limit of the n-stage problem. One of the advantages of this approach is that it enables us to determine, for denumerable χ, a necessary and sufficient condition for the existence of a bounded solution to a functional equation which characterizes the optimal policy.

2. Stationary Deterministic Optimal Policies

The following theorem was originally proven by Derman [2] for the special case that χ is denumerable. The following proof is new; it makes use of a technique used by Taylor [5].

Theorem 1: If there exists a bounded Baire function f(x), $x \epsilon \chi$ and a constant g, such that

$$(1) \qquad g + f(x) = \min_{k \epsilon A} \{C(x,k) + \int_{y \epsilon \chi} f(y) \; dP(y|x,k)\} \quad x \epsilon \chi$$

then there exists a stationary deterministic policy R* such that

$$g = \phi(x,R^*) = \min_{R} \phi(x,R) \text{ for all } x \epsilon \chi$$

and R* is any policy which, for each x, prescribes an action which minimizes the right side of (1).

Proof: For any policy R

$$E_R \{ \sum_{t=1}^{n} [f(X_t) - E_R(f(X_t)|S_{t-1})] \} = 0$$

But

$$E_R[f(X_t)|S_{t-1}] = \int_{y\epsilon\mathcal{X}} f(y)dP(y|X_{t-1},\Delta_{t-1})$$

$$= C(X_{t-1},\Delta_{t-1}) + \int_{y\epsilon\mathcal{X}} f(y)dP(y|X_{t-1},\Delta_{t-1}) - C(X_{t-1},\Delta_{t-1})$$

$$\geq \min_{k\epsilon A} \{C(X_{t-1},k) + \int_{y\epsilon\mathcal{X}} f(y)dP(y|X_{t-1},k)\} - C(X_{t-1},\Delta_{t-1})$$

$$\geq g + f(X_{t-1}) - C(X_{t-1},\Delta_{t-1})$$

with equality for R* since R* is defined to take the minimizing action. Hence

$$0 \leq E_R\{ \sum_{t=1}^{n} [f(X_t) - g - f(X_{t-1}) + C(X_{t-1},\Delta_{t-1})]\}$$

or

$$(2) \qquad g \leq E_R \frac{f(X_n)}{n} - E_R \frac{f(X_0)}{n} + E_R \frac{\sum_{t=1}^{n} C(X_{t-1},\Delta_{t-1})}{n}$$

with equality for R*. Letting $n \to \infty$ and using the fact that f is bounded, we have that $g \leq \phi(R,X_0)$ with equality for R*, and for all possible values of X_0. QED.

Remark: Note that the above proof doesn't make use of the fact that A is finite or that C(x,k) is bounded.

Let $g_n(x)$, $n = 1,2,\ldots$ satisfy

$$(3) \qquad g_1(x) = \min_k C(x,k)$$

$$g_{n+1}(x) = \min_k \{C(x,k) + \int_{y\epsilon X} g_n(y)dP(y|x,k)\}$$

Note that $g_n(x) = \min_R \sum_{t=0}^{n-1} E_R[C(X_t,\Delta_t)|X_0 = x]$. The following corollary was proven by Derman [2] for the denumerable case.

Corollary 1: Under the conditions of theorem 1, there is a M such that

$$|g_n(x) - ng| < M \quad \text{for all } n,x$$

Proof: Let M' be such that $|f(x)| < M'$. By (2) we have that $g \leq 2M' + g_n(x)$. Again from (2), by letting $R = R^*$ we have that $g \geq g_n(x) - 2M'$. QED.

Fix some state - call it 0 - and let

$$(4) \qquad f_n(x) = g_n(x) - g_n(0) \qquad \text{all } n,x$$

One has from (3) that

$$(5) \qquad g_{n+1}(0) - g_n(0) + f_n(x) = \min_k \{C(x,k) + \int_{y\epsilon X} f_n(y)dP(y|x,k)\}$$

We shall now determine sufficient (and in the denumerable case necessary and sufficient) conditions for the existence of a bounded Baire function f(x) and a constant g satisfying (1).

Theorem 2: If $\{f_n\}$ is a uniformly bounded equicontinuous family of functions then

(i) there exists a bounded continuous function f(x) and a constant g satisfying (1).

(ii) $\lim_{n\to\infty} (g_{n+1}(x) - g_n(x)) = g$ for all $x\varepsilon\chi$.

Proof: By the Ascoli Theorem there exists a subsequence $\{f_{n_k}\}$ and a continuous function f such that $f_{n_k}(x) \to f(x)$. Now $g_{n+1}(0) - g_n(0)$ is bounded (since costs are bounded) and so we can also require that $g_{n_k+1}(0) - g_{n_k}(0) \to g$. Hence by (5) and the bounded convergence theorem we have that $g + f(x) = \min \{C(x,k) + \int_{y\varepsilon\chi} f(y)dP(y|x,k)\}$.

For any subsequence $\{n'\}$ of $\{n\}$ there is a sub-subsequence $\{n''\}$ such that $\lim(g_{n''+1}(0) - g_{n''}(0))$ exists. By the above this limit must be g. Thus $g = \lim_n (g_{n+1}(0) - g_n(0))$. The result follows since 0 is any arbitrary state. QED.

If χ is denumerable, then $\{f_n\}$ can always be taken to be equicontinuous by considering the discrete topology. We thus have

Corollary 2: If χ is denumerable, then a necessary and sufficient condition for the existence of a bounded function f(x) and constant g satisfying (1) is that there is a $M < \infty$ such that $|g_n(x) - g_n(0)| < M$ for all n,x.

Proof: Sufficiency follows from the above theorem and necessity follows from Corollary 1. QED.

For any policy R, $\beta\varepsilon(0,1)$, let $\psi(x,\beta,R) = \sum_{t=0}^{\infty} \beta^t E_R[C(X_t,\Delta_t)|X_0 = x]$.

A policy R_β such that $\psi(x,\beta,R_\beta) = \min_R \psi(x,\beta,R)$ for all $x\varepsilon\chi$ is said to be β-optimal.

We shall need the following result given by Blackwell [1]:
If A is finite, and $C(\cdot,\cdot)$ is bounded then, for each $\beta\varepsilon(0,1)$, there is a stationary deterministic policy R_β which is β-optimal. Furthermore $\psi(x,\beta,R_\beta)$ is the unique solution to

$$(6)\qquad \psi(x,\beta,R_\beta) = \min_{k\in A} \{C(x,k) + \beta \int_{y\in X} \psi(y,\beta,R_\beta)dP(y|x,k)\}$$

and any policy which, when in state x, selects an action which minimizes the right side of (6) is β-optimal.

Fix some state - call it 0 - and let

$$(7)\qquad f_\beta(x) = \psi(x,\beta,R_\beta) - \psi(0,\beta,R_\beta)$$

then

$$(8)\qquad g_\beta + f_\beta(x) = \min_{k} \{C(x,k) + \beta \int_{y\in X} f_\beta(y)dP(y|x,k)\}$$

where

$$g_\beta = (1-\beta)\ \psi\ (0,\beta,R_\beta)$$

In analagous fashion to Theorem 2 we have

Theorem 3: If $\{f_\beta\}$ is a uniformly bounded equicontinuous family of functions then

(i) there exists a bounded continuous function f(x) and a constant g satisfying (1).

(ii) $(1-\beta)V_\beta(x) \to g$ as $\beta \to 1^-$ for all $x\in X$.

Proof: Same as proof of Theorem 2.

For any stationary deterministic policy R let x(R) be the action .. chooses when in state x. We say that $\lim_n R_n = R$ if, for each x, there exists $N_x < \infty$ such that $x(R_n) = x(R)$ for all $n \geq N_x$.

The following was proven in [4] for denumerable X. The proof for arbitrary X is identical

Theorem 4: Under the conditions of Theorem 3

(i) for some sequence $\beta_r \to 1^-$, $R^* = \lim_r R_{\beta_r}$

(ii) if $R = \lim_r R_{\beta_r}$, where $\beta_r \to 1^-$ then R is optimal - i.e.

$\phi(x,R) = g$ for all $x \in \mathcal{X}$.

The following two conditions were given by Taylor [5] to prove equicontinuity of $\{f_\beta\}$ in the special case of a replacement process:

(a) For every $k \in A$, $C(\cdot,k)$ is continuous.

(b) For every $x \in \mathcal{X}$, $k \in A$, $P(|x,k)$ is absolutely continuous with respect to some σ-finite measure μ on B and it possesses a density $p(y|x,k)$ also assumed to be a Baire function in x. Furthermore, for every $x \in \mathcal{X}$, $k \in A$

$$\lim_{x' \to x} \int |p(y|x,k) - p(y|x',k)| d\mu(y) = 0$$

Theorem 5: If conditions (a) and (b) are satisfied then

(i) $|f_\beta(x)| < M$ for all $x,\beta \Rightarrow \{f_\beta\}$ is equicontinuous

(ii) $|f_n(x)| < M$ for all $x,n \Rightarrow \{f_n\}$ is equicontinuous

Proof: Follows directly from (5) and (8) and conditions (a), (b).

A sufficient condition for the uniform boundedness of $\{f_\beta\}$ is given in [4].

3. Reduction of Average Cost Case to Discounted Cost Case

We shall need the following assumption

Assumption (I): There is a state - call it 0 - and $\alpha > 0$, such that $P\{X_{t+1} = 0 | X_t = x, \Delta_t = k\} \geq \alpha$ for all $x \in \mathcal{X}$, $k \in A$.

For any process satisfying the above Assumption consider a new process with identical state and action spaces, with identical costs, but with transition probabilities now given by

$$P'(B|x,k) = \begin{cases} \dfrac{P(B|x,k)}{1-\alpha} & \text{for } 0 \notin B \\ \dfrac{P(B|x,k) - \alpha}{1-\alpha} & \text{for } 0 \in B \end{cases} \qquad B \in \mathcal{B}$$

Let $\psi'(x,\beta,R)$ be the total expected β-discounted cost, and let R'_β be the β-optimal policy, all with respect to the new process. Letting $f'(x) = \psi'(x,1-\alpha,R'_{1-\alpha}) - \psi'(0,1-\alpha,R'_{1-\alpha})$ we have by (8) that

$$(9) \quad \alpha\psi'(0,1-\alpha,R) + f'(x) = \min_k \{C(x,k) + (1-\alpha) \int_{y\in\mathcal{X}} f'(y)dP'(y|x,k)\}$$

$$= \min_k \{C(x,k) + \int_{y\in\mathcal{X}} f'(y)dP(y|x,k)\}$$

And thus the conditions of Theorem 1 are satisfied. It follows that $g = \alpha\psi'(0,1-\alpha,R'_{1-\alpha})$ and the optimal average-cost policy is the one which selects the actions which minimize the right side of (9). But it is easily seen that $R'_{1-\alpha}$ does exactly this. Hence the optimal average cost policy is precisely the $1-\alpha$-optimal policy with respect to the new process; and the optimal expected average cost per unit time is $\alpha\psi'(0,1-\alpha,R'_{1-\alpha})$.

The above result was proven in [4] for the denumerable case by showing that $\phi(x,R) = \alpha\psi'(0,1-\alpha,R)$ for any stationary policy R.

This result also holds for arbitrary χ. However this in itself does not show that $R'_{1-\alpha}$ is optimal. (It does in the denumerable case because it can be shown that Assumption (I) implies that $\{f_\beta\}$ is uniformly bounded and thus by Theorem 3 there exists a stationary deterministic policy which is optimal.)

4. Concluding Remarks

Results given in [4] which dealt with ε-optimal policies and replacement processes (Sections 3 and 4) carry over to the more general spaces χ considered here. The proofs are identical (with integrals replacing sums in the obvious places).

REFERENCES

[1] Blackwell, David (1965), Discounted Dynamic Programming, Annals of Mathematical Statistics, 36, 226-235.

[2] Derman, Cyrus (1966), Denumerable State Markovian Decision Processes - Average Cost Criterion, Annals of Mathematical Statistics, 37, 1545-1554.

[3] Derman, Cyrus and Lieberman, Gerald J. (1966), A Markovian Decision Model for a Joint Replacement and Stocking Problem, Management Science, 13, 609-617.

[4] Ross, Sheldon M. (1967), Non-Discounted Denumerable Markovian Decision Models, To appear in Annals of Mathematical Statistics.

[5] Taylor, Howard (1965), Markovian Sequential Replacement Processes, Annals of Mathematical Statistics, 36, 1677-1694.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA · R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Stanford University Department of Statistics Stanford, California	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

Arbitrary State Markovian Decision Processes

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report

5. AUTHOR(S) (Last name, first name, initial)
Sheldon M. Ross

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
January 8, 1968	12	5

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Nonr-225(53) b. PROJECT NO. NR-042-002 c. d.	105 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Logistics and Mathematical Statistics Br. Office of Naval Research Washington, D.C. 20360

13. ABSTRACT

Arbitrary state, finite action Markovian decision processes are studied with respect to the (long-run) average cost criterion. The problem is treated both as a limiting case of the discounted cost problem and also as a limit of the n-stage problem. Sufficient conditions are given for the existence of an optimal rule and for it to be of stationary deterministic type.

DD FORM 1 JAN 64 1473

UNCLASSIFIED
Security Classification

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Markovian Decision Process						
Arbitrary State Space						
Stationary Deterministic Optimal Rule						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.

DD FORM 1 JAN 64 1473 (BACK)

UNCLASSIFIED

Security Classification